DOCUMENT RESUME

ED 395 990 TM 025 184

AUTHOR Sireci, Stephen G.
TITLE Problems and Issues in Linking Assessment across Languages.
PUB DATE 27 Oct 95
NOTE 20p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bilingualism; *Educational Assessment; *Equated Scores; Intercultural Communication; *Item Response Theory; *Second Languages; *Test Construction; Test Format; Translation
IDENTIFIERS Experts; *Language Minorities; Linking Metrics

ABSTRACT

Test developers continue to struggle with the technical and logistical problems inherent in assessing achievement across different languages. Many testing programs offer separate language versions of a test to evaluate the achievement of examinees in different language groups. However, comparisons of individuals who took different language versions of a test are not valid unless the score scales for the different versions are linked or equated. This paper discusses the psychometric problems involved in cross-lingual assessment, reviews linking models that have been proposed to enhance score comparability, and provides suggestions for developing and evaluating a model for linking different language versions of a test. The review of the literature did not reveal a linking model that completely resolved the problem, but it suggested that three designs are superior to methods that employ translation only or that use expert judgment to certify score equivalence. These designs are: (1) item response theory (IRT) linking using separate monolingual groups; (2) IRT linking via a bilingual group; and (3) matched monolingual group designs. (Contains 4 figures and 43 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 395 990

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
STEPHEN G. SIRECI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM
In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view. FL

Problems and Issues in Linking Assessments Across Languages

Stephen G. Sireci[1]

University of Massachusetts at Amherst

Paper presented at the Annual Conference of the Northeastern Educational Research Association, Ellenville, NY, October 27, 1995

---

[1]The quality of this paper was enhanced by recent conversations on this topic with Ron Hambleton and H. Swaminathan, and by suggestions from Liane Patsula who reviewed an earlier draft.

BEST COPY AVAILABLE

TM025184

Abstract

Test developers continue to struggle with the technical and logistical problems inherent in assessing achievement across different languages. Many testing programs offer separate language versions of a test to evaluate the achievement of examinees in different language groups. However, comparisons of individuals who took different language versions of a test are not valid unless the score scales for the different versions are linked or equated. This paper discusses the psychometric problems involved in cross-lingual assessment, reviews linking models that have been proposed to enhance score comparability, and provides suggestions for developing and evaluating a model for linking different language versions of a test.

"If English was good enough for Jesus, it's good enough for the schoolchildren of Texas."
Texas Governor James "Pa" Ferguson, 1917, after vetoing bill to finance the teaching of foreign languages in the classroom.

INTRODUCTION

There is a growing need to measure achievement in a variety of languages. This need stems in large part from the increasing number of students throughout the U.S. who are not proficient in English, and the desire to compare the educational achievement of students in different countries. Comparing achievement across different languages poses formidable challenges for test developers because unintended differences in the content or difficulty between different language versions of a test may account for observed differences between groups or individuals. This paper explores the psychometric problems inherent in cross-lingual assessment, reviews previous research in this area, and presents suggestions for enhancing the comparability of scores derived from different language (DL) versions of a test.

Recently, several substantial contributions have been added to the vast literature on methods and standards for cross-lingual assessment. Hambleton (1993) for example, discussed problems inherent in translating (adapting) tests across languages and addressed the critical validity issues associated with this process. Many of these problems and issues are addressed in the *Guidelines for Adopting Educational and Psychological Tests* forthcoming from the International Test Commission (ITC, in press; largely summarized by Hambleton, 1994), and in Geisinger's (1994) treatise on cross-cultural normative assessment.

Before attempting to link DL tests onto a common scale, it must be demonstrated that the constructs measured by the DL tests are comparable. Such demonstration has been termed construct equivalence or content parallelism. This paper does not address the construct equivalence of DL tests; rather it focuses on the problem of linking DL tests designed to measure constructs that are generalizable across different language groups. It is also assumed here that the test context and item formats are appropriate for the DL groups. For elaborate discussions of evaluating construct equivalence across languages, see Geisinger (1992, 1994), Hambleton (1993, 1994), Hui and Triandis (1985), Martin and Berberoglu (1991), and Olmedo (1981).

A fundamental theme of the recent writings in the area of test adaptation is that merely translating a test from one language to another does not signify score comparability across languages (Angoff & Cook, 1988; Geisinger, 1994; Hambleton, 1993; Prieto, 1992). Rather, scores resulting from DL tests must be statistically adjusted for meaningful comparisons to be made. The preferred form of adjustment is to link the DL tests onto a common score scale. Linking DL versions of a test onto a common scale is required to accurately separate effects stemming from true differences in proficiency between DL groups from differences due to the separate language versions of the test. Some contemporary examples of linking assessments in cross-lingual research include:

- comparison of the educational achievement of students in different countries, who receive instruction in different languages (International Association for the Evaluation of

Educational Achievement (IEA), 1994; LaPointe, Mead, & Phillips, 1989; Miura, Okamoto, Kim, Steere, & Fayol, 1993),

- evaluation of the cross-cultural generalizability of attitudes or psychological constructs (Ellis, 1989; Hulin, Drasgow, & Komocar, 1982; Hulin & Mayer, 1986; Martin and Berberoglu 1991), and

- evaluation of the academic proficiency of non-English speaking students in the United States with respect to their English-speaking peers (Angoff & Cook, 1988; CTB, 1988; O'Brien, 1992).

In the United States, linking DL tests onto a common scale is also relevant in personnel, licensure, and industrial competency testing where persons of different language backgrounds are tested with respect to a job or content domain not related to English language proficiency (e.g., Ramos, 1981). Most linking studies in the U.S. have focused on linking tests translated into Spanish to the original English-language version. However, the linking problem is generic across languages. In Israel, for example, the Psychometric Entrance Test (required for entrance into Israeli universities) is linked across six different languages (Beller, 1994).

METHODS USED TO LINK TESTS ACROSS LANGUAGES

Attempts to link different language versions of a test onto a common scale can be classified into three general research design categories: 1) separate monolingual group designs, 2) bilingual group designs, and 3) matched monolingual group designs. In the separate monolingual group design, source- and target-language versions of a test are separately administered to source- and target-language examinee groups. Items considered to be equivalent across the source- and target-language versions of the test (anchor items) are used to link the DL tests onto a common score scale. In the bilingual group design, a group of examinees who are proficient in both the source and target languages is used to link the separate monolingual groups. The matched monolingual group design attempts to create equivalent source- and target-language groups by matching examinees in the two groups on criteria deemed relevant to the proficiency measured, or uses observed differences on the matching criteria to adjust for differences in test or item performance. These designs have been used in various forms and combinations. The most popular and praised methods use item response theory (IRT) models to calibrate the DL tests onto a common scale. A review of these designs reveals their strengths, limitations, and underlying assumptions.

IRT Linking Using Separate Monolingual Groups

IRT models have been used in a variety of settings to link DL tests. Educational applications include Angoff and Cook's (1988) linking of the Scholastic Aptitude Test to its Spanish counterpart the Prueba de Aptitud Académica, and O'Brien's (1992) and Woodcock and Muñoz-Sandoval's (1993) linking of English and Spanish language proficiency tests. Examples from

industrial testing include the linking of the English and Hebrew, and English and Spanish versions of the Job Descriptive Index (Hulin & Mayer, 1986; and Hulin, Drasgow, & Komocar, 1982). Examples are also found in psychological testing, such as Ellis (1989) who investigated linking English and German intelligence tests, and Martin & Berberoglu (1992) who linked English and Turkish versions of a social desirability scale. These applications all used a unidimensional IRT model to calibrate the DL tests; however the particular model used varied from one study to another.

Although there are variations in the procedures used in these studies, linking DL tests using two DL monolingual groups and IRT typically involves the following steps:

1) The source language (e.g., English language) test is translated into the target language (e.g., Spanish language) via a comprehensive series of adaptation techniques (see Hambleton, 1993, 1994).

2) The source-language test is administered to source-language examinees, and the target-language test is administered to target language examinees.

3) The source- and target-language tests are separately calibrated using an IRT model.

4) A scale transformation procedure (e.g., Stocking & Lord, 1983) is used to place the item parameter estimates for the DL tests onto a common scale. The target-language test item parameters are usually transformed to the source-language test scale.

5) Translated items are evaluated for invariance across the DL tests. IRT-based methods for evaluating differential item functioning (DIF) are typically used to determine item equivalence across languages (e.g., the IRT chi-square technique; Lord, 1980). The DIF evaluation procedure may be iterative, where items that initially display DIF are eliminated from the subsequent stratifying variable (e.g., "purifying" θ).

6) Items considered invariant across the DL tests are used as anchor items to calibrate the tests onto a common scale. Items that are not statistically equivalent across the tests are either deleted or considered unique to the separate language versions. The anchor-item equating procedure could be IRT-based (e.g., concurrent calibration constraining anchor item parameters to be equal), or could be based on a classical anchor-item design.

These general steps do not apply to all studies that used IRT to link DL tests, but are characteristic of the general approach. The Angoff and Cook (1988) study went beyond these general steps by first pre-testing items in English and Spanish populations. This preliminary step allowed them to identify items that appeared statistically equivalent in both populations. The equivalence was re-evaluated with the subsequent calibration sample.

Linking DL tests via IRT modeling is a preferred strategy because of the sample invariance

properties of IRT (i.e., item parameters estimated using IRT are not dependent on the specific range or level of proficiency of the examinee sample used; Hambleton, Swaminathan, & Rogers, 1991). Because DL groups may differ markedly with respect to the proficiency measured, it is necessary to account for these differences when linking tests. However, a criticism of using IRT models to link DL tests is that the sample invariance properties of IRT may not hold over samples derived from DL examinee groups.

Assumptions Underlying the Monolingual IRT Approach

An evaluation of the assumptions underlying the monolingual IRT approach for linking DL tests reveals the controversy surrounding item parameter invariance across DL groups. When DL tests are separately calibrated in each language group, the only assumption required for IRT calibration is that the test items are measuring an essentially unidimensional construct. However, more restrictive assumptions are required for calibrating these two separate tests onto a common scale. Linking the DL tests requires: construct equivalence across languages, unidimensionality of the pool of DL items, and common items across both tests. This last requirement is the most difficult to realize in practice, and in some cases, it is difficult to determine whether it has been accomplished at all.

As an illustration of this predicament, consider the monolingual IRT approach outlined above. Without anchor items between the DL tests, it is not possible to link the tests onto a common scale. Concurrent calibration does not form a common scale because differences in proficiency not accounted for by the model would affect the item parameter estimates for the original and translated items. Because only source language examinees take the source language items, the parameters for these items are referenced only to the source language group. Likewise, the target language item parameters are referenced to only the target language examinee group. The sample invariance properties of IRT models may not extend to these DL samples because it is not clear whether the two DL groups represent samples from a single population, or samples from different populations.

The problem of uncertainty of ability differences between groups is easily solved using common anchor items between test forms. Anchor items, by definition, are equivalent in both forms of a test that are to be linked. However, with DL tests, determination of anchor items is problematic. It is clear that translated items cannot be considered equivalent without empirical evidence (Angoff & Cook, 1988; Hambleton, 1993). But to provide empirical evidence of item invariance across languages, a valid matching criterion is required. Thus a criticism of the IRT monolingual linking method is that when translated items are evaluated for DIF across the DL groups, they are not on a common scale necessary for evaluation of DIF. The IRT proficiency scale (θ–scale) is a fallible matching criterion because there are no true common items. Scale transformation procedures, such as the Stocking-Lord procedure do not resolve this dilemma, as they require anchor items or some other means for accounting for differences in proficiency between the separate calibration groups.

As an example of the potential confound between test differences due to lack of comparability of translated test forms and differences between DL group proficiencies, consider two language groups who, on average, differ one-half of a standard deviation unit with respect to the proficiency measured. To make the example more concrete, assume that we are trying to link English- and Spanish-language versions of a multiple-choice science achievement test for junior high school students across U.S. English-speaking students, and Spanish-speaking students in Costa Rica. Let us assume further that the distribution of science proficiency is the same for the two populations with the exception of the center of the distribution: the Costa Rican distribution centers at $\theta = .5$, while the U.S. distribution centers at $\theta = 0$. To link the tests we utilize a monolingual group design using the three-parameter logisitc IRT model. Given this hypothetical "true" difference in science proficiency between these two groups, translated items with true difficulty parameter differences as large as .5 may appear equivalent if they are calibrated concurrently, or if they are transformed to be on a common scale using a transformation procedure that does not account for the difference in group proficiencies.

This predicament is illustrated in Figures 1, 2, and 3. Figure 1 illustrates the hypothetical distribution of science proficiency for these two groups on the hypothetical ("true") English-Spanish scale (θ_T). Figure 2 presents the ICCs for an original and translated item, where the items have different location (difficulty) parameters. Because the true, common, θ-scale accounts for the differences in proficiencies between these two groups, comparing the ICCs illustrates that the item does not function equivalently across the two languages. Obviously, the adaptation of the item from English to Spanish made the item harder. Figure 3 illustrates how the ICCs would appear if they were scaled concurrently (or transformed onto a common scale) without accounting for the group differences in science proficiency (θ_O is the theta scale estimated from the observed responses). The ICCs in Figure 3 look identical.

Thus the major drawback of the separate monolingual group IRT approach is the inability to separate the DL group proficiency differences from differences due to the DL tests (or items) themselves. Theoretically, the monolingual groups IRT method can be effective only when the equivalence of the anchor items can be defended outside of the IRT calibration model.

If it can be demonstrated that a sufficient number of anchor items are truly equivalent across DL tests, then the parameters for these items can be constrained to be equal across the tests when calibrating the tests concurrently, or can be used to form the common scale via the Stocking-Lord procedure. Anchor items that do not require translation or adaptation (that are identical in both the source and target languages) do not need to be evaluated for statistical equivalence across the DL tests. Demonstration of item equivalence across languages is more feasible when dealing with constructs that are language independent, such as computational items in mathematics. Although largely unexplored in the literature on linking DL tests, the use of non-verbal items to form a set of anchor items could also be used to link DL tests, or to form a common scale for evaluating cross-lingual DIF of the remaining items.

Figure 1: Hypothetical Proficiency Distributions
("True" Common Scale)

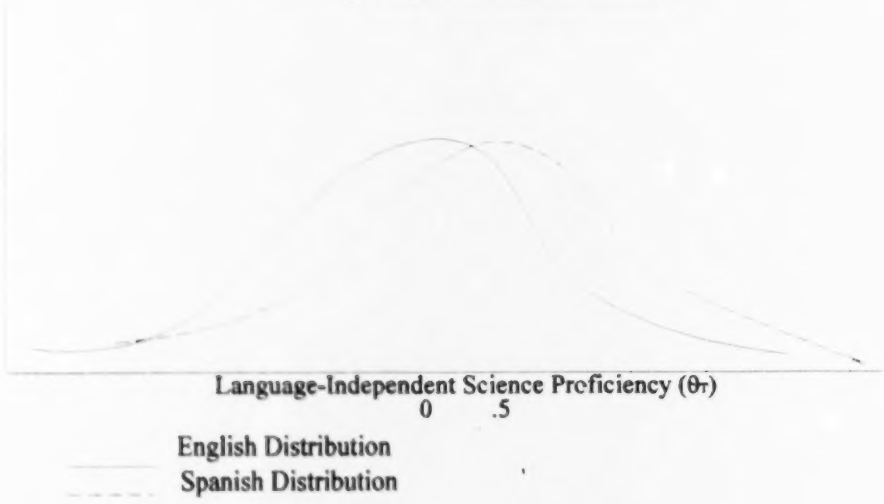


Figure 2: Original & Translated Item
On Hypothetical Common Scale

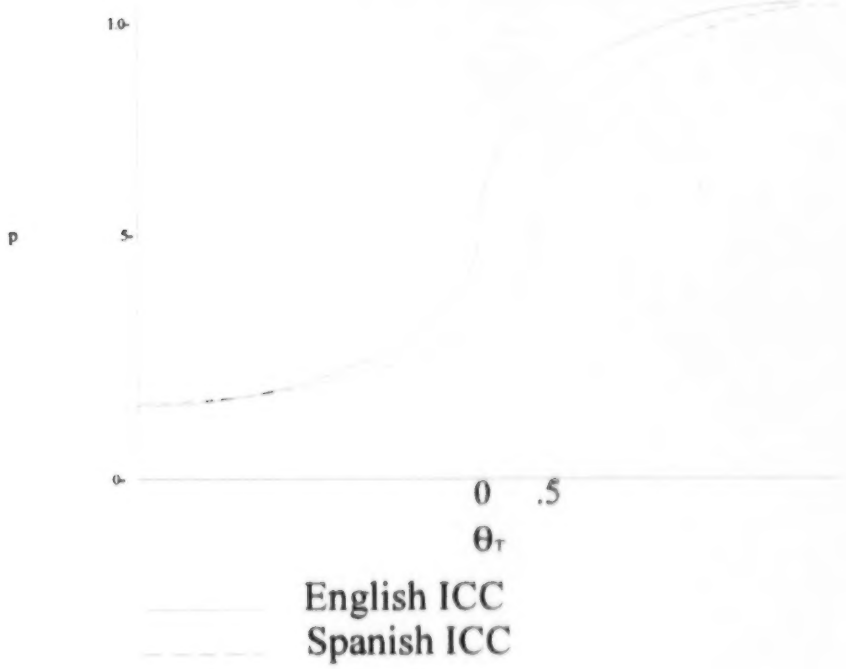
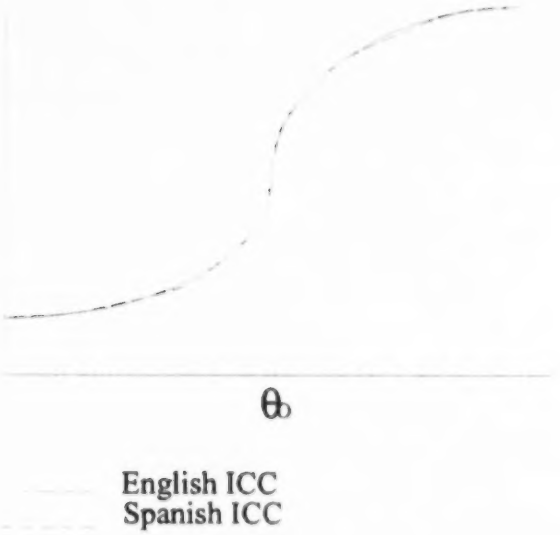


Figure 3: Concurrently-Calibrated ICCs



In the absence of non-linguistic anchor items, the validity of the monolingual group IRT approach cannot be fully evaluated. However, even though Angoff and Cook (1988) did not employ non-linguistic anchor items, they were able to interpret the prevalence of DIF across languages in terms of the content areas measured by the items. Items more closely associated with linguistic features displayed DIF more often. Far more verbal items displayed cross-lingual DIF, and the analogy items, which were considered the most context-laden, exhibited the highest level of DIF. Very few mathematics items exhibited cross-lingual DIF. These findings suggest intuitively that the "common" metric used to evaluate item parameter invariance across languages was effective. Thus the example portrayed in Figures 1 through 3, and the associated criticism of the monolingual groups IRT method, may arise only when the item adaptation procedures produce relatively few comparable items. The item adaptation procedures used by Angoff and Cook were comprehensive. It may be that adherence to strict test adaptation guidelines (e.g., Hambleton, 1993; 1994) provides a sufficient number of invariant items for the formation of a common scale for DIF analysis.

Additional problems in calibrating DL tests using separate monolingual groups are non-overlapping portions of the ability distributions for the separate DL groups, and differences between the variance of these distributions. If the DL proficiency distributions overlap only partially, then anchor item equivalence may be possible for only a portion of the θ-distribution for both groups (i.e., only for the interval of overlap). If this problem occurs, then the anchor items used to link the DL tests would not fully represent the distribution of operational items. Non-representative anchor tests used in anchor-item equating designs have been shown to bias equating results (Cook & Petersen, 1987; Klein & Jarjoura, 1985).

IRT Linking via a Bilingual Group

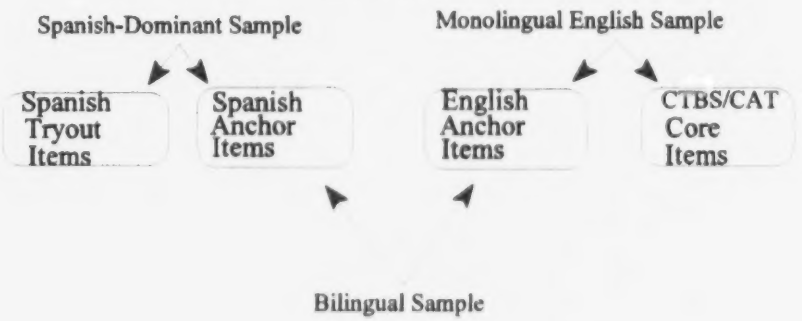
One method utilized to separate the effects of group differences across languages from the effects of differences due to the DL tests, is to use a group of examinees who are proficient in both source and target languages (Boldt, 1969). Theoretically, a group of bilingual examinees would be equally proficient in both languages with respect to the proficiency measured. Therefore, group differences in proficiency are eliminated. Given this assumption, simultaneous calibration of translated items onto a common scale can be conducted, and the DL items can be evaluated for DIF. Non-DIF items serve as anchor items to link the DL tests onto a common scale. If the test translation produced items that were truly equivalent in both languages, the probability of a bilingual examinee getting an item correct in either the source or target language would be equal. If the probability of success on an item differed between the source and target language, then the item could not be considered invariant across languages.

There are three potential variants of the bilingual group design. The most common is the single-group design where a single bilingual group takes both language versions of the test (or sets of potential anchor items) in counterbalanced order. This design maximizes language group

comparability, but may be affected by a practice effect from taking two tests designed to be identical except for language medium. A second option is to use two randomly equivalent bilingual groups, each of whom takes one language form of the test. This design avoids practice effects, but compromises group comparability. The third option is to use two randomly equivalent bilingual groups who respond to unique source and target items (i.e., items that are not translations of one another). This design is currently being explored by Berberoglu and his colleagues using two separate test forms comprising items alternating between the source and target language (G. Berberoglu, personal communication, September 30, 1995).

A comprehensive example of the bilingual group linking design is the method used to link the Spanish Assessment of Basic Education (SABE) to the Comprehensive Test of Basic Skills (CTBS) and the California Achievement Tests (CAT; CTB, 1988). In this study, students who were English-Spanish bilingual responded to pilot sets of Spanish and English anchor items. These items were written to measure the same skills and content areas. The English anchor items were also administered to a monolingual English group and the Spanish anchor items were administered to a monolingual Spanish group. This research design is depicted in Figure 4. The performance of the bilingual group on the pilot anchor items was used to select a set of final anchor items that functioned similarly in both their English and Spanish versions.

Figure 4: Schematic of SABE Research Design
(from CTB, 1988, p. 6)



Although the bilingual group approach directly addresses the problem of disentangling group differences from test differences, it has several major drawbacks. First, it is very difficult to find a group of examinees that are "equally proficient" in two languages. Even if such a group were found, equivalent language proficiency may not signify equal proficiency in both languages with

respect to the attribute measured. Second, a bilingual group of examinees may represent only a small portion of the proficiency scale of one or both of the monolingual groups for whom the tests are designed. Bilingual examinees are different from their monolingual counterparts with respect to language proficiency, and so it is likely they will not be representative of either language group with respect to the proficiency measured. Thus the bilingual group approach for linking DL tests has both practical and technical limitations; namely, problems in identifying truly bilingual examinees, and the lack of generalizability of the results from bilingual examinees to their monolingual cohorts. However, a bilingual group could be useful for evaluating the invariance of anchor items used in monolingual linking designs.

Matched Monolingual Group Designs

The matched monolingual group linking design attempts to control for group differences in proficiency by matching examinees on criteria deemed relevant to the proficiency measured, rather than by accounting for group differences via anchor items. Two approaches can be used: creation of equivalent groups by selecting pairs of examinees in DL groups with similar values on the matching criteria, or differences between groups on the criteria are used to account for group differences in the proficiency measured. Caliper matching and matching using propensity scores (Rindskopf, 1986; Rosenbaum & Rubin, 1983) are applicable to this problem. Caliper matching refers to matching on score intervals rather than on exact criterion values. Propensity scores refer to scores that describe "the conditional probability of assignment to a particular treatment given an observed vector of covariates" (Rosenbaum & Rubin, 1983; p.41).

There are not many examples of the matched monolingual group linking design, probably due to the obvious problem of finding relevant and available matching criteria. Tamayo (1990) matched 120 students age 8 to 16 on age, sex, school, grade, and academic achievement (as estimated by their teachers) before evaluating translation differences of the WISC-R vocabulary subtest (32 vocabulary items). Although this approach employed a matched-groups design, it essentially sought out to prove the null hypothesis (i.e., no difference between translated versions of the test) using a relatively small sample, and so the efficacy of this design needs further exploration. A further disadvantage of the matched group design is that the validity of the matching criteria must be established, and it must be equivalent in both language populations.

Although the matched-groups linking design has not received a great deal of attention in cross-lingual linking studies, matching examinees in DL groups could reduce the effect of group proficiency differences that threaten the validity of the separate monolingual group anchor item design. The effects of matching on equating parallel forms of a test written in the same language have been investigated, but the results are equivocal (Kolen, 1990; Skaggs, 1990). Cook, Eignor, and Schmitt (1989), Eignor, Stocking, and Cook, (1990), and Livingston, Dorans, and Wright (1990) did not find improvement over non-matched designs, while Wright and Dorans (1993) concluded that matching did improve equating results. Wright and Dorans, and Livingston et al., suggested that equating may be improved via matching on propensity scores, but thus far, propensity scores have not been applied to the equating problem.

BEST COPY AVAILABLE

CONCLUSIONS

The preceding analysis of models used to link DL tests provides more questions than answers regarding valid cross-lingual assessment. Given the current trend toward cross-national educational comparisons (e.g., Feuer & Fulton, 1994; IEA, 1994), it is clear that ignorance of linguistic factors affecting such comparative studies is unacceptable. It is also clear that accounting for these factors poses formidable challenges for cross-lingual educational researchers.

Suggestions for Future Cross-Lingual Linking Research

The review of the literature did not reveal a linking model that completely resolved the problem of linking tests across languages. Of course, it is always easier to point our weaknesses in previous research than it is to provide suggestions for improvement. However, it is not the intention of this paper to draw a pessimistic picture of techniques for linking tests across languages. Although the methods reviewed have shortcomings, they go far beyond the assumption that scores derived from DL tests are directly comparable. These state-of-the-art techniques represent considerable progress from the earlier days of cross-cultural research where differences in test content across languages were not even considered as potential confounds affecting observed group differences (Brislin, 1970; Prieto, 1992). As Hambleton (1994) pointed out

> The common error is to be rather casual about the test adaptation process, and then interpret the score differences among samples or populations as if they were real. This mindless disregard of test adaptation problems and the need to validate instruments in the cultures where they are used has seriously undermined the results from many cross-cultural studies. (p. 242)

The three designs reviewed in this paper are far superior methods for promoting score comparability across DL tests than are methods that employ translation only, or that use "expert" judgment to certify score equivalence.

Given the strengths and weaknesses of current methods for linking DL tests, the following suggestions for future research are proposed:

1) Test developers should construct non-verbal items, or items minimally associated with language skills, to be used to form the anchor between DL tests. The goal is to defend the equivalence of items across the DL tests in a manner independent of the calibration model. The current trend in performance-based assessment could facilitate such item development. For example, on a science test, examinees could be asked to identify elements in the periodic table with specific properties (e.g., 3 electrons), or be asked to complete an unfinished drawing illustrating the flow of magnetic forces. The use of non-verbal assessment has been successful in cross-cultural research where language-independent behavioral observation is employed. Shelley-Sireci, Fracasso, Busch-Rossnagel, and Lamb (1995) for example, illustrated the stability of mother-infant interaction styles across English and Latino populations using a language-independent assessment of interaction behavior.

2) When test translation is employed, the procedures should be governed strictly and conform to the guidelines promoted by the ITC (Hambleton, 1994). Adherence to rigid guidelines will facilitate linking of the tests on a common scale. Inequalities introduced in the translation process will at best reduce the item pool, and at worst, bias the calibration.

3) Although it is probably not defendable to use a "bilingual" group as the primary link between DL tests, bilingual examinees could be used to screen potential anchor items; especially those containing linguistic context. Due to the "representation" problem, the bilingual group cannot be used to validate anchor item equivalence, but they could be used to identify items that are clearly not equivalent across languages.

4) Future research should explore matching DL monolingual examinees to tease out the effects of language-group proficiency differences from differences due to the test translation process. Matching via propensity scores is theoretically appealing, but has not been evaluated with respect to linking DL tests. As with the bilingual group design, matching DL groups will probably not result in a defendable linking design in its own right, but may be useful for supplementing designs using separate monolingual groups.

5) Multidimensional IRT models (e.g., Ackerman, 1994) should also be explored for linking DL tests. If separate dimensions can be identified for source or target language proficiency, and the proficiency measured by the test, then the latter dimension can be used as a "purified" matching criterion for evaluating DIF among original and translated items.

These proposals are suggested for enhancing score comparability of DL tests. Only empirical research can determine their utility. Regardless of the linking design employed, linking tests across languages is only one component of cross-lingual research involving DL groups. In particular, questions of construct and predictive validity must also be studied (Anastasi, 1992; Geisinger, 1992; 1994; Hambleton, 1993; 1994). However, when test score-based inferences focus on comparing the proficiencies of DL examinees, adjustment for differences due to the measurement procedure (i.e., linking) is requisite. Contrary to the opinions of "Pa" Ferguson, ignoring the importance of multiple languages in a global society limits the validity of contemporary educational research.

References

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 7, 255-278.

Anastasi, A. Introductory remarks. In K.F. Geisinger (Ed.) Psychological Testing of Hispanics (pp. 1-7). Washington, DC: American Psychological Association.

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report No. 88-2). New York, NY: College Entrance Examination Board.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli Universities. Educational Measurement: Issues and Practice, 13, 12-20.

Boldt, R.F. (1969). Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers. College Entrance Examination Board Research and Development Report 68-69, No. 3, Princeton, NJ: Educational Testing Service.

Brislin, R.W. (1970). Back-translation for cross-cultural research. Journal of Cross-cultural psychology, 1, 185-216.

CTB (1988). Spanish assessment of basic education. Technical report. Monterey, CA: McGraw Hill.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1989). Equating Achievement tests using samples matched on ability. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Cook, L.L. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of the effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. Applied Measurement in Education, 3, 37-55.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-920.

Ellis, B.B., & Kimmel, H.D. (1992). Identification of unique cultural response patterns by means of item response theory. Journal of Applied Psychology, 77, 177-184.

Feuer, M.J., & Fulton, K (1994). Educational testing abroad and lessons for the United States. Educational Measurement: Issues and Practice, 13, 31-39.

Geisinger, K.F. (1992). Fairness and selected psychometric issues in the psychological testing of Hispanics. In K.F. Geisinger (Ed.) Psychological Testing of Hispanics (pp. 17-42). Washington, DC: American Psychological Association.

Geisinger, K.F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304-312.

Hambleton, R. K. (1993). Translating Achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: a progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hui, C.H., & Triandis, H.C., (1985). Measurement in cross-cultural psychology: a review and comparison of studies. Journal of Cross-Cultural Psychology, 16, 131-152.

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology, 67, 818-825.

Hulin, C.L., & Mayer, L.J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71, 83-94.

International Association for the Evaluation of Educational Achievement (1994). TIMSS main study manuals: population 1 and 2. Hamburg: Author.

International Test Commission (in press). Guidelines for adapting test instruments and establishing score equivalence. European Journal of Psychological Assessment.

Klein, L.W., & Jarjoura, D. (1985). Effect of number of common items in common-item equating with nonrandom groups. Journal of Educational Measurement, 22, 197-206.

Kolen, M. J. (1990). Does matching in an equating work? A discussion. Applied Measurement in Education, 3, 97-104.

LaPointe, A.E., Mead, N.A., & Phillips, G.W. (1989). A world of differences: an international assessment of mathematics and science (Report No. 19-CAEP-01). Princeton, NJ:

Educational Testing Service

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? Applied Measurement in Education, 3, 73-95.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Martin, M.R., & Berberoglu, G. (1991). Initial efforts in construct validation for the Turkish Marlowe-Crowne Social Desirability Scale. In B. Thompson (Ed.), Advances in educational research: substantive findings, methodological developments (pp. 25-36). Greenwich, CT: JAI Press.

Martin, M.R., & Berberoglu, G. (1992). Further construct validation of the Turkish Marlow-Crowne Social Desirability Scale using the Rasch model. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX, February 1.

Miura, I.T., Okamoto, Y., Kim, C.C., Steere, M., & Fayol, M. (1993). First graders' cognitive representation of number and understanding of place value: cross-national comparison -- France, Japan, Korea, Sweden, and the United States. Journal of Educational Psychology, 85, 24-30.

O'Brien, M.L. (1992). A Rasch approach to scaling issues in testing Hispanics. In K.F. Geisinger (Ed.) Psychological Testing of Hispanics (pp. 43-54). Washington, DC: American Psychological Association.

Olmedo, E.L. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.

Prieto, A. J. (1992). A method for translation of instruments to other languages. Adult Education Quarterly, 43, 1-14.

Rindskopf, D. (1986). New developments in selection modeling for quasi-experimentation. In W. M. K. Trochim (Ed.) Advances in quasi-experiential design and analysis, no. 31. San Francisco, CA: Jossey-Bass.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education, 3, 53-71.

Shelley-Sireci, L.M., Fracasso, M.P., Busch-Rossnagel, & Lamb, M.E. (1995). Mother-

infant social and instrumental interaction in culturally diverse populations. Paper presented at the annual meeting of the American Psychological Society, July, New York, NY.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. Applied Measurement in Education, 3, 105-113.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Tamayo, J.M. (1990). A validated translation into Spanish of the WISC-R vocabulary subtest words. Educational and Psychological Measurement, 50, 915-921.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. European Journal of Psychological Assessment, 3, 1-16.

Wright, N. K., & Dorans, N. J. (1993). Using the selection variable for matching or equating. (Research Rep. No. RR-93-4). Princeton, NJ: Educational Testing Service.